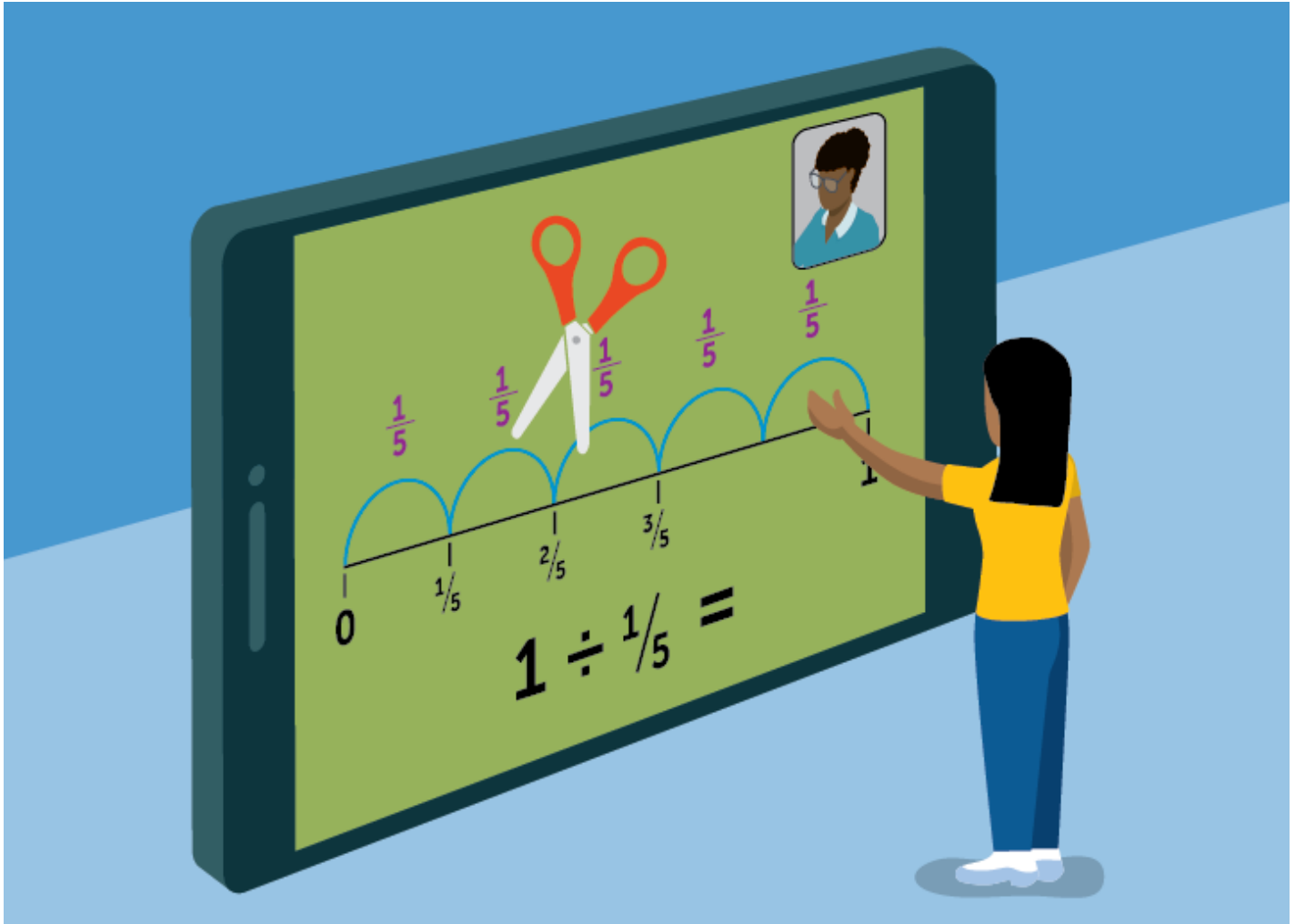


Transfer for Future Learning of Fractions within Cognition's Microtutoring Approach

Jeremy Roschelle, Britte Haugan Cheng, Nicola Hodkowski

Lina Haldar, and Julie Neisler¹

April 29, 2020



¹ Authors are Digital Promise staff except Britte Haugan Cheng (MenloEDU) & Lina Haldar (LCHaldar Consulting)

Suggested Citation

Roschelle, J, Cheng, B. H, Hodkowski, N., Haldar, L., and Neisler, J. (2020). *Transfer for future learning of fractions within Cognition's microtutoring approach [Project Report]*. San Mateo, CA: Digital Promise. Retrieved from <http://hdl.handle.net/20.500.12265/95>

Acknowledgements

This research was supported by grants from the Bill & Melinda Gates Foundation, the Chan-Zuckerberg Initiative, and Schmidt Futures, under a subcontract from Cognition. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funders. We thank Elizabeth Tipton, Associate Professor in Statistics at Northwestern University, for graciously providing meta-analytic advice. The Digital Promise Global and MenloEDU team thanks the Cognition team for their collaboration in conducting this research, expert panelists, the tutors and all project contributors.

Contact Information

Email: jroschelle@digitalpromise.org

Digital Promise:

Washington, DC:

1001 Connecticut Avenue NW, Suite 935
Washington, DC 20036

San Mateo, CA:

2955 Campus Dr. Suite 110
San Mateo, CA 94403

Website: <https://digitalpromise.org/>

Executive Summary

In this exploratory research project, we designed and began to validate a measurement approach that could provide indication of a student's ability to transfer their mathematics understanding to future, more advanced mathematical topics. Assessing transfer of learning in mathematics and other topics is an enduring challenge. We sought to invent and validate an approach to transfer that would leverage Cognition's use of online 1:1 tutoring and would be relevant to their future product development.

The scientific basis for our process of invention had three components. Our first scientific basis was in **research on mathematics learning**. There is a long and strong history of research on fractions as part of the larger learning trajectory in rational number, multiplicative reasoning and representational competencies (like the number line). We focused on relatively near transfer as most appropriate for our student population. Second, we applied **Evidence-Centered Design** as an overarching assessment design and development framework. Third, we were inspired by the work of Dan Schwartz and team on **Preparation for Future Learning (PFL)**. Because we did not strictly follow PFL precedents, we termed our approach "Transfer for Future Learning" (TFL). PFL inspired TFL in that we looked for transfer into a future mathematical topic in the same learning progression.

In the first phase of work, we specified potential TFL tasks through a process of domain analysis and domain modeling, which mapped related topics that could be conceptually linked via targeted instruction. An expert panel reviewed this work and guided how we proceeded. In the second phase of work, we designed, fielded, and improved two potential TFL tasks, one we refer to as "Fraction Division" and the other as "Fraction Generalization." The testing occurred through online 1:1 tutoring sessions, which functioned as cognitive labs, and led to many improvements. The resulting tasks have a three-part structure:

Portion A: Elicit student's prior knowledge

Portion B: Instruction to promote transfer to a new topic

Portion C: Assess performance on the new topic via transfer items

At the end of this process of iterative improvement, we noted both ways in which the tasks were ready for further testing and elements that still could be improved. In a third phase, we worked with a population of students who had just completed approximately 18 sessions of 1:1 online tutoring on relevant fraction topics that precede those addressed in TFL tasks, "Fraction Division" and "Fraction Generalization."

In the third phase of work, we gathered and analyzed data on the TFL tasks more systematically. This phase was conducted immediately after Cognition finished their fractions tutoring with students and gave a fractions posttest (the Cognition posttest). Our hypothesis was "Students with higher scores on a Cognition posttest should be better able to progress in TFL tasks." Participants were fifth-grade students in one school that participated in Cognition's tutoring study. Participating students met with their tutor for two additional 30-minute sessions, one for each TFL task. Tutors were trained on how to administer these tasks. We

received anonymized pretest and posttest data from Cognition (the data from their larger experiment). We also received data about the tutoring sessions and the use of the “Fog Stone Isle” game. We analyzed a data set containing 22 audio and video recordings for each task. Two independent raters scored the recordings for correct math answers and features related to explanation and representation (after achieving suitable interrater reliability).

To examine our hypotheses, we first analyzed the correlation between a student’s Cognition posttest score and score for the overall TFL task. We found correlations of 0.68 (Fraction Division) and 0.57 (Fraction Generalization), both of which were statistically significant. We also found correlations from the “prior knowledge” of each TFL task (portion A) to the later portion where students transfer the prior knowledge to a new topic (portion C). The finding of a statistically significant correlation between the Cognition posttest and our overall TFL task confirms our broad hypothesis: students who have stronger fraction knowledge at the end of the Cognition tutoring ought to be better able to transfer that knowledge to more advanced mathematics. The correlations between the Cognition posttest and portion C of the TFL tasks offer a fine-grained view of the same underlying trend. We also looked at how hard the TFL tasks were for students and found that although some students earned high TFL task scores, other students found transfer to be challenging, which was to be expected.

There are limitations to this study. First, due to logistical and recording quality issues beyond our control, we had a small data set. Hence, this work should be replicated or extended with a larger group of students. Second, we had anticipated testing for transfer with students who were already above a threshold for prior knowledge, whereas in the actual data, some students were still showing weaker prior knowledge. The work should be replicated or extended with large numbers of students who are ready for transfer. Third, we observed factors which could contribute to noisy data and which could be addressed through further design refinement.

Cognition reported that the assessment design process and findings were helpful. The early design phase of the work clarified the expected learning progress for Cognition’s tutoring and what kinds of tutoring decisions could have later transfer implications. In the second phase, during the agile development of the TFL tasks, we discovered difficulties eliciting student knowledge, instructional issues in the tutoring process, and user interface challenges in video conferencing software. Our findings were informative to Cognition’s improvements in these areas. Finally, as Cognition plans to teach “Fraction Division” and “Fraction Generalization” in more extended tutoring work, the TFL tasks provide useful guidance on how to connect prior knowledge to these advanced and difficult topics.

For the field of assessment, the approach taken in this project was innovative in how it used online tutoring sessions to gather information on each student’s knowledge and abilities with regard to transferring from recent learning to future topics. The validation data collected is promising yet preliminary. Further work is needed to refine the assessment designs and to provide further validity evidence.

Introduction: Rationale and Goals

In this exploratory research project, our team's goal was to design and begin validation of a measurement approach that could provide indication of a student's ability to transfer their mathematics knowledge to future, more advanced mathematical topics. Assessing transfer of learning in mathematics and other topics is an enduring challenge (Evans, 1999; National Resource Council, 2000). Although the field is making progress in assessing transfer in mathematics (Johnson, McClintock, & Hornbein, 2018; Lobato 2009, 2012; Rayner, Bernard, & Osana, 2013; Rittle-Johnson, Loehr, & Durkin, 2017), the process of developing assessments of transfer is slow and expensive. Our study explored the potential for an agile, iterative process that could help R&D teams who are designing technologies that promote transfer of learning.

The one-year exploratory research project was conducted under a subcontract within a larger Cognition project. The larger project developed Cognition's 1:1 online tutoring approach ("microtutoring") and collected a first round of data on it with students who were struggling with fraction concepts. Our rationale for doing this exploratory research in the Cognition context was that (a) Cognition emphasizes conceptual understanding (b) conceptual understanding is necessary to and supportive of transfer (Barnett & Ceci, 2002) (c) transfer will not be measured by a typical unit test or posttest and (d) Cognition wanted to use information on transfer to improve their approach.

One important distinction that we paid attention to throughout the entire process was between **validating** a measure and **evaluating** Cognition's approach. In the long term, the payoff of Transfer for Future Learning work will be in the ability to evaluate student learning that emphasizes conceptual understanding. In essence, transfer is a powerful crucible in which to examine the payoff of an approach that develops students' conceptual understanding. However, one cannot use a measure to evaluate a product before the measure has been validated. Therefore, we recommend against interpreting this report as an evaluation of Cognition's approach.

In this one-year project, our emphasis was on initial design and validation of a measure. The work was framed as exploratory research because transfer is notoriously difficult to measure. Indeed, there is no standard practice in assessment design and development that is known to produce sound transfer measures for mathematical content like this. Our effort was neither routine nor predictable. For example, initially we had planned for game-based delivery and we had assumed that we would have considerable control of how TFL tasks would be presented in the game. However, as the work emerged, we transitioned to tutor-delivered tasks, which had major implications for all the research that followed. We used the context of Cognition's microtutoring to invent a measurement approach and then conducted early stage validation research on the invention.

As we worked, three contributions of this work were top of mind as potential outcomes. In what ways could exploratory research on transfer...

1. provide insights to Cognition that were immediately useful for their product development efforts?
2. inform the assessment field about promising approaches for developing measures of transfer within content-specific learning trajectories?
3. illuminate further assessment research and development that would need to be done to design and develop a transfer measure that could be fully validated?

In our discussion, we will report on these questions. In addition, we coordinated with a Mathematica evaluation of Cognition's approach and other non-Cognition projects in a portfolio of related investments. In Appendix 1, we provide data on metrics that were defined to support Mathematica's evaluation approach.

Phase 1: Laying the Groundwork (February - May 2019)

The activities in this phase were (a) to analyze the domains of mathematical knowledge relevant to our transfer goal and to Cognition's content, (b) to develop initial patterns for an assessment, and (c) to get feedback from a panel of experts on both the domain analysis and design patterns. Here we summarize a report that was delivered in June 2019 on this phase (Roschelle, Cheng, & Cohen, 2019).

Theoretical Framework

Our theoretical framework draws on three literatures. First, in mathematics, student understanding of fractions predicts future learning of mathematics up to, through, and beyond eighth- or ninth-grade Algebra (Booth & Newton, 2012; Siegler et al., 2012). There are several plausible reasons for the empirical relationship. Fractions are one of the earliest curricular topics that engage substantial symbolic, conceptual, and computational complexity. Fractions also lead directly into topics of rational number and multiplicative reasoning and beyond (DeWolf, Bassock, & Holyoak, 2015; Empson, Levi, & Carpenter, 2011; Hackenberg, 2013; McMullen et al., 2015; Saxe, Diakow, & Gearhart, 2013; Thompson & Saldanha, 2003). Further, all these concepts are used in solving algebra problems, along with related representations (number lines) and practices (generalizing).

Second, Evidence-Centered Design (ECD) is a theoretically sound, well-accepted process for design, development, and validation of high-quality assessments (Mislevy, Steinberg, & Almond, 1999). ECD organizes assessment R&D into component processes that include domain analysis, domain modeling, the creation of design patterns for assessments, and building operational assessments and scoring systems. We draw from ECD a focus on defining tasks, evidence, and rubrics that align to yield insight on students' knowledge, skills, and abilities.

Third, we were inspired by the Preparation for Future Learning (PFL) approach (Bransford & Schwartz, 1999; Chin et al., 2019; Schwartz & Martin, 2004). Like PFL, we focus on transfer from current understanding to future concept—transfer within a learning progression. Also, like PFL, we provide instructional support (“bridging”) to reduce the difficulty of transfer; students rarely transfer knowledge alone. We adopt a three-phase protocol: (1) eliciting prior knowledge (2) bridging instruction from prior knowledge to a target concept and (3) performance on the target concept. We call our approach “Transfer for Future Learning” (TFL) because we diverge from PFL in some details.

Domain Analysis

The goal of the domain analysis activity was to define the mathematical content of instruction in the Cognition experience and the related mathematical content later in a learning progression. To analyze the domains of mathematical knowledge we communicated extensively with Cognition about which fraction concepts they cover and how they teach these concepts to students. We learned that fraction equivalence, comparison, and addition would likely be covered with students. We learned about key visual representations that Cognition would include, like area and number line models. We also noted Cognition’s conceptual approaches, such as using fractions in the context of measurement and focusing on the meaning of numerator and denominator. We examined mathematical standards and curricula to determine what these topics might transfer to in future mathematics. Broadly, the topics fit into the development of rational number and multiplicative reasoning, which takes place over 3-5 years of mathematics instruction and includes topics of ratios, proportions, linear functions, and graphing.

Domain Modeling

Next, we moved to domain modeling, which specified transfer opportunities. We did this by examining both the presumed strengths of students’ learning with Cognition and the difficulty of transfer opportunities in the domain analysis. We realized that many of these topics in the domain analysis would be too far of a stretch for struggling students, who had recently come to terms with fractions. Thus, we chose three initial topics:

1. Generalizing comparisons of fraction magnitude
2. Going from measurement whole number division to (initial progress on) fraction division
3. Moving from the use of the number line for fractions towards the number line for both positive and negative fractions

For each topic, we elaborated considerable detail on what a TFL task might look like.

Platform Considerations

We also discussed potential platforms for the assessment with Cognition. Initially we thought that the TFL tasks would be presented in “Fog Stone Isle,” Cognition’s computer platform. But,

increasingly, Cognition chose to emphasize microtutoring by expert human tutors via an online video conference call. Together, we realized that delivering the tasks in “Fog Stone Isle” might require development work that was a distraction from Cognition’s overall plan. Therefore, we began to contemplate a shift from the game platform to the microtutoring platform.

Expert Panel Review

After this preliminary work, we met with three outside experts for a day to review this work. The experts were highly engaged in the meeting. They specifically appreciated the following aspects of Cognition’s approach and this exploratory research project:

- Challenging content area of fractions
- Formative information and analytics from the microtutoring sessions
- Evidence-Centered Design approach
- Combination of human and technology in microtutoring sessions

With Mathematica, we had agreed to report on two pre-registered metrics (see Appendix 1), each of which were on a scale of ok, good, or great. The experts rated our domain analysis. Since they did not rate it great, they shared their thoughts on how to make it great. After much discussion, the experts rated our design patterns as good (Fraction Generalization) and as okay (Fraction Division and a third design pattern). In the expert’s view, the third pattern, regarding the number line for negative numbers, was in need of the most work.

Wrapping Up Phase 1

We followed the recommendations of the expert panel. We fixed the domain analysis and decided to develop two tasks, “Fraction Division” and “Fraction Generalization.” We subsequently met with Cognition and agreed to change plans. We decided not to deliver the assessment in the game; instead we would deliver it via microtutoring.

Phase 2: Agile Development (June 2019 - February 2020)

In the agile development process, we designed two TFL tasks, tested them with a series of cohorts of students, and iteratively improved them. Both tasks used the same overall three-part structure:

Portion A: Prior Knowledge. The first activity elicits relevant prior knowledge (the source for transfer). This activity also provides information on whether students are ready for transfer (e.g. do they show conceptual understanding of the expected existing knowledge?).

Portion B: Instruction. An instructional activity introduces the student on how to transfer the concept; it provides a bridge between the source and the target of transfer. The tutor can be active during this instruction.

Portion C: Transfer. The transfer activity reveals whether the student can use the prior concept in the new way (the target for transfer) and how they use conceptual understanding in doing so.

We first describe the final version of the two tasks, then the process of iteration from initial to final form. Each of the two tasks focused on different types of conceptual transfer needed for proficiency in mathematics. The “Fraction Division” task focused on transfer to a new concept of division with fractions and whole numbers. The “Fraction Generalization” task was on transfer to an understanding and recognizing of patterns in existing knowledge as a way of further conceptualizing the magnitude of fractions (in any situation).

The Fraction Division TFL Task

The “Fraction Division” task focused on students’ understanding of quotitive division (Tzur et al., 2013) with whole numbers and moved towards transfer to quotitive division of a whole number divided by a fraction. (Quotitive division means dividing the measurement of a whole into an unknown number of groups with a known amount per group or unit measurement.) Along with students’ understanding of quotitive division, it used the concept of iterating a fractional unit as the source for transfer. This required students to take a total amount, and an amount to count by (amount per group) in order to find the total amount of groups required. An example follows:

- **Transfer from:** There are 6 inches of string, how many 2-inch pieces can you make?
- **Transfer to:** There are 2 inches of string, how many $\frac{1}{2}$ inch pieces can you make?

The Fraction Generalization TFL Task

The “Fraction Generalization” task focused on students’ understanding of fraction magnitude and moved towards transfer to generalizing this understanding. Students were guided to use what they know about a denominator or a numerator of certain fractions and then tell about the pattern they saw in that certain group of numbers. An example follows:

- **Transfer from:** Place $\frac{2}{5}$ and $\frac{4}{5}$ on the number line. Explain which one is greater and why
- **Transfer to:** Explain if the expression is never, sometimes, or always true for values of n : $\frac{1}{n} < \frac{2}{n}$

Iterative Testing Process

For each TFL task, an iteration of testing consisted of:

1. Finalizing the task for testing
2. Training microtutors to implement the task
3. Collecting data as microtutors used the task with a cohort of students
4. Analyzing the data
5. Discussing what improvements to make

Data was collected in the form of field notes on microtutoring sessions that were directly observed as they occurred or observations from recordings of microtutoring sessions. The basic analysis template was to score each student response as right or wrong, to note the quality of student explanations, and report additional notes on the format and process of the session.

Extent of Iteration

The table below reports the extent of iteration that occurred. We iterated more on the Fraction Division task in response to our sense that it needed more work.

Month of iteration	Fraction Division Task	Fraction Generalization Task
1) June	4 students	4 students
2) June	4 students	4 students
3) October	3 students	(skip)
4) November	2 students	3 students
5) December	4 students	4 students
6) early January	3 students	(skip)
7) late January	4 students	(skip)
Totals	7 iterations, 24 students	4 iterations, 15 students

Table 1: Number of students involved in testing during each iteration of a TFL task

Issues Observed During Iterations

We kept notes on the issues observed in each iteration and potential design fixes that might be tried. We summarize those notes below:

- **Student difficulty understanding language.** Sometimes our prompts were too generic to elicit further explanation or there was too much to read. We streamlined and simplified the language.
- **Supporting student explanation.** We added prompts for tutors to ask students to elaborate explanations or responses in an effort to get clearer or more extended explanations.

- **Tutor interventions.** We noticed that tutors diverged from our instructions or expectations, for example, providing hints or guidance on the performance tasks. We enhanced instructions and tutor training.
- **Sessions too long.** Throughout iterations, we found that sessions took a longer time than expected. We reduced the number of subtasks and simplified complex reading or requests. We gave tutors more guidance on how long to spend on each subtask.
- **(*) Student difficulty with representations.** Students sometimes had difficulties with number lines, especially when asked to draw them on their own. We added labels and tick marks to number lines, and sought to reduce the need for students to draw their own number lines.
- **(*) Tutor understanding of mathematics.** We sometimes noticed tutors introducing a mathematical concept (fractions as part of a whole) that were less well suited to the intended transfer than our desired approach (fractions as a measure represented on a number line). We also wondered about how each tutor's instructional approach during regular tutoring related to the number line. As mathematical issues came up throughout the agile development process, we discussed with Cognition so they could consider adjustments to their overall guidance to tutors.
- **(*) Tutoring environment limitations.** We noticed that it was sometimes hard for students or the tutor to draw in the environment, so we tried to reduce the amount of drawing. In addition, the task structure had to fit the presentation requirements of the environment, which sometimes required rethinking what was presented on each screen.

Of these difficulties, we felt more confident that we had addressed the first four (without “*”) issues by the end of the iterations. We were somewhat less confident that we had addressed the last three. Regarding the ECD underpinnings of our work, we also considered which of these would be “construct relevant” or “construct irrelevant” variance. Notably, we would categorize student difficulty with representations as “construct relevant”—that is, this is yielding meaningful data about students’ understanding. We would categorize Tutor Understanding of Mathematics and Tutor Environment Limitations as “construct irrelevant”—that is, these are distracting from our ability to measure the target student understandings.

Metrics

Before the beginning of Phase 1, we pre-registered a metric with Mathematica that had two aspects: how many untested revisions we made for each task and the degree to which students completed each task during the last round of testing. See Appendix 1 for details. Due to the changes in our exploratory research plan over the course of the year, these became less meaningful or relevant to our research. In particular, these metrics made more sense when we were planning for a game-based, not a tutor-delivered task—the metrics do not contemplate the role of the tutor or the tutoring environment and both of these proved to be critical variables. We also would frame any consideration of student “completion”

differently in a tutor-delivered task than in a game-based task as the nature of time constraints that emerge in each delivery situation are quite different.

Reflections

Overall, our team had frequent and robust discussions of our degree of satisfaction with the tasks as we finished agile development. Our key thoughts were as follows:

1. We were satisfied with the three-part structure of the TFL task. We were satisfied in particular, that the structure of the task strengthened the quality of the measurement of student learning.
2. We recognized the degree to which variation between tutors and constraints of the tutoring environment impacted the TFL task was greater than we had anticipated, leading to construct irrelevant variance, which was undesirable.
3. We had ongoing concerns about students' existing knowledge and readiness to use the number line; while this impacted both tasks, it seemed to impact the "Fraction Division" task more. We also wondered how tutors related to the number line. As we discuss below, this is a "construct relevant" source of variance; it is highly relevant to Cognition's approach and to how the field develops fraction concepts in a way that builds to future learning.

The three-part structure of each TFL task (prior knowledge, instruction, transfer) did not change during the iterations. We found it very helpful to include portion A as this often revealed problems in students' prior knowledge; if we had only the transfer performance (portion C), we would not know if weak performance was a matter of missing conceptual foundations or weak transfer. Overall, we found that transfer performance was challenging for many students and yet we also saw some students succeed. The instruction during portion B did not make transfer too easy for students; indeed, this opportunity to learn how to transfer prior knowledge appeared to be essential to transfer. Overall, the three portions worked well together.

When we had planned for game-based delivery, we had assumed that we would have considerable control of how TLF tasks would be presented in the game. Hence, our focus was initially on variation in how students reacted to tasks. However, we quickly found that there were two other important sources of variation in the tutor-delivered tasks. Tutors differed in how they presented the TFL tasks (see issues, above). Tutors could modulate how long each problem within the tasks took, and sometimes this led to very long sessions. In addition, tutors often had differing ways of conceptualizing the math and this could influence students in the tutoring tasks. The environment in which the tasks were provided also made it hard to work with some representations. For example, we couldn't directly spin a spinner to show choice of a specific number for a variable. It was hard for tutors and students to physically draw number lines, and then partition and iterate on those lines. Overall, whereas we had planned for 100% focus on students as a source of variation, we ended up spending many iterations working on issues related to tutors and the environment in which the task was provided. We hastened to add that there were benefits to switching to the tutor-

delivered tasks (vs. game-delivered). For example, it seemed the supportive environment of tutoring made students willing to try to do something new and difficult; it may have made exploring transfer more comfortable. Also, explanations are easier to elicit when there is someone to talk to, and when successful at getting students to explain it was very informative.

Finally, based on the domain analysis, we had expected a number line representation to be something that students were gaining mastery on through their use of “Fog Stone Isle” and their engagement in microtutoring. To some extent, we came across student number line issues due to timing; we interacted with students before they had much microtutoring on number lines. But we also saw students struggle with number lines (and be more comfortable with other strategies for thinking about fractions). Drawing a number line was hard for students, both conceptually and practically in the environment, so we provided pre-drawn number lines and focused on using them to explain. We sometimes wondered the degree to which the number line was something that tutors had incorporated in their own conceptual approach. We minimized the dependence of the “Fraction Generalization” task on the number line representation, and continued to iterate on the “Fraction Division” task to try to simplify number line use (and to anticipate additional number line tutoring that would occur between the time of our testing and March, the time of major data collection). This issue remained an uncertainty as we wrapped up development of the tasks. We note that it is a productive uncertainty—a number line representation of fraction understanding is central to all modern treatment of fractions (e.g. in the Common Core and other state standards) and it is important for preparation for future learning, as number lines continue to be used throughout middle school. Hence, learning more about how tutors and students work with to show understanding via number lines is substantively important to Cognition’s further product development. We planned to learn more in the next phase.

Wrap Up to Phase 2

We decided that both TFL tasks were good enough to merit testing in the next phase of the exploratory research. Nonetheless, we resolved to consider the issues we had encountered when reporting the next phase as potential limitations or potential explanations for any patterns observed. In other words, we believe the tasks each had merits with regard to detecting student conceptual understanding and student transfer (the tasks detect “construct relevant variance”), and that we were now well aware of key sources of noise that might arise (we were aware of “construct irrelevant variance” to watch out for in further testing).

Phase 3: Transfer Study (February 2020 - April 2020)

To further validate the two TFL tasks, we arranged with Cognition to gather data from students who had participated in their tutoring program. Our primary overall hypothesis for this study was:

Students with higher scores following Cognition tutoring should progress further in TFL tasks than those with lower scores.

Given our hypothesis that there will be a positive, linear relationship between scores following Cognition tutoring and TFL task scores, the null hypothesis would be that there is no relationship between scores following Cognition tutoring and TFL task scores. If the anticipated relationship does not exist, future work would be needed to understand why these two measures showed no relationship. If the anticipated relationship *does* exist, this constitutes evidence that there is, indeed, a relationship between demonstration of knowledge through Cognition tutoring and our measure, which seeks to bridge that knowledge to new learning.

We also expected to see variation on the TFL task as transfer is difficult for students. Fraction division is notoriously difficult to conceptualize—most instruction simply focuses on invert and multiply only. Fraction generalization involves a use of variables that fifth-grade students are not likely to be strong in. In particular, the Common Core State Standards introduce letter as an unknown in third grade, however letter as a variable is not introduced until sixth grade. We did not expect to see uniformly high scores. However, if some students were able to transfer successfully, this would be consistent with our expectations for a tutorial program that emphasizes conceptual understanding.

Participants

Participants (fifth-grade students) for the transfer study came from two classrooms in one school that participated in Cognition’s overall tutoring study. Other sites were not able to participate due to timing issues. We obtained consent from the parents for students to participate in the transfer study. We only involved students who had tutoring in the Cognition study because a precondition for the transfer study was that the student had a relationship with a tutor. Conversely, if we included students who were meeting a tutor for the first time, there would be an obvious confound between their transfer experience and their newness to online tutoring.

Task and Training Procedure

Participating students met with their tutor for two additional 30-minute sessions, one for each TFL task. The TFL tasks were the tasks resulting from Phase 2.

Tutors were trained about one week before administering the TFL tasks. The training lasted for roughly an hour and fifteen minutes. The key information covered during the training included the following for each task:

- Goals for tasks
- Structure of tasks
- Time considerations
- Discussion activities for each task to sample how to promote student explanations

It is important to note that the training did not include any material on the mathematical knowledge and the concepts of each of the TFL tasks. We consider this to be a limitation of this study (see Limitation section).

Data Collection and Scoring

We received anonymized pretest and posttest data from Cognition (the data from their larger experiment). We also received data about the tutoring sessions and the use of the “Fog Stone Isle” game.

We recorded the audio and video from the tutoring sessions. Some sessions had poor quality recordings (either audio or visual) and were dropped from analysis. This resulted in $n=22$ sessions for “Fraction Division” and the same number of sessions for “Fraction Generalization” (with some differences in which students were in each group, because recordings were not consistently unusable for any one student).

To score the data, we achieved scoring agreement for each of the TFL tasks between two raters. To do so, we began with the “Fraction Generalization” task and created an initial rubric for scoring both students’ correctness and explanation for three prior knowledge questions and four transfer questions. Each rater then individually watched and scored three “Fraction Generalization” task sessions (7 correctness scores and 7 explanation scores per session, total of 42 scores). Following the individual scoring, the two raters, a statistician, and an additional member of the team met to calibrate scores. Based on the discussions, adjustments were made to the rubric. Each rater then individually reconciled the same three task sessions (a total of 72 scores). Based on the reconciliation, another meeting took place with the two raters, the statistician, and an additional team member. They resolved all disagreements and a final rubric was created. Each rater then individually scored 3 additional sessions and the raters met to calibrate on those scores (a total of 93 scores). Since, this final meeting included very few scoring disagreements (<5), the rater team moved to individual scoring of the remaining 16 “Fraction Generalization” sessions. One rater scored nine sessions, while the other scored seven sessions. In total, raters scored 22 “Fraction Generalization” task sessions, 6 sessions together and individually split scoring 16 sessions, a total of 682 scores between the prior knowledge questions and transfer questions.

The raters then moved to the “Fraction Division” task scoring. Similarly, a rubric was created prior to scoring any sessions to measure students’ correctness and explanation for three prior knowledge questions and four transfer questions. Each rater then individually scored three “Fraction Division” sessions (a total of 72 scores). The two raters met to discuss minor disagreements and revise the rubric as needed. Following, each rater scored three additional sessions (a total of 72 scores). The two raters met again to resolve any disagreements (<5) and make final revisions to the rubric. The raters then moved on to score the remaining 16 sessions. One rater scored nine sessions, while the other scored seven sessions. In total, raters scored 22 Fraction Division task sessions, 6 sessions together and individually split scoring 16 sessions, a total of 528 scores between the prior knowledge questions and transfer questions.

Quantitative Analysis Plan

To test our hypothesis (above), we planned several comparisons for each TFL task. We conducted one confirmatory validation, which was between the scores on both parts of task and the Cognition posttest.

The data and analysis for this hypothesis align with a metric that was pre-registered with Mathematica (Appendix 1).

We also conducted three exploratory analyses to interrogate components of that overall result in more detail:

1. Between the score on only the prior knowledge part (portion A) of the task and the Cognition posttest
2. Between the score on only the transfer part (portion C) of the task and the Cognition posttest
3. Between the two parts of the task, transfer and prior knowledge (portions C and A).

In each case, we plotted scatter plots and computed a correlation value and a test of statistical significance.

In addition, we explored descriptive statistics of students game-play in the “Fog Stone Isle” (FSI) game. This data included:

- Number of minutes of active play, defined as the number of minutes directly attributed to a specific domain within FSI, not simply being logged into the game;
- Number of sessions, where session end points were determined by inactivity for 10 minutes, then the session is considered ended 10 minutes earlier. If activity should resume after the 10-minute threshold, the first new action is considered the start of a new session; and
- Specific data from within each of the learning domains, including introduction to fractions, adding fractions, equivalent fractions, dividing fractions, adding decimals, multiplying fractions, and fraction as ratios.

Exploratory Qualitative Analysis Plan

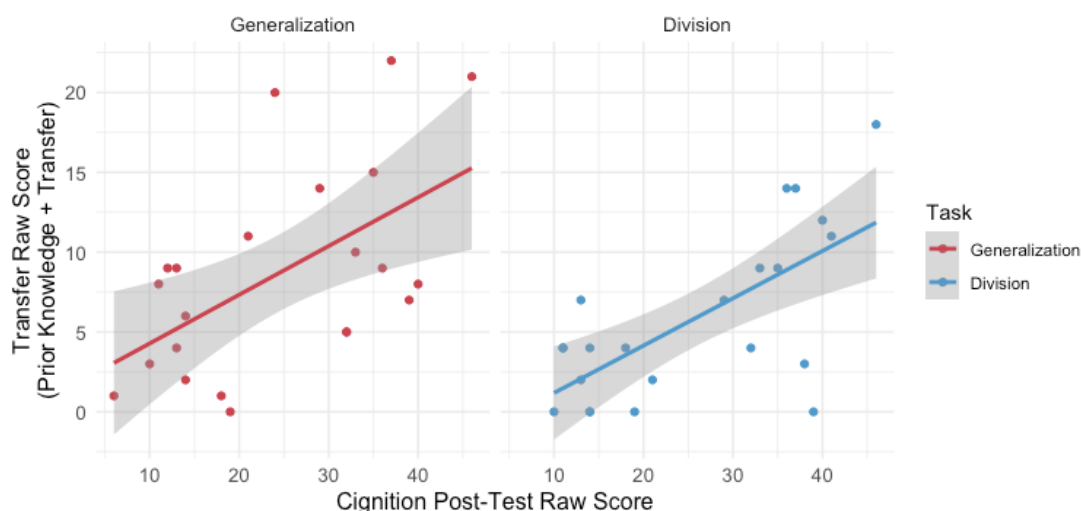
We also looked at a subset of five recordings for insights on how the content of the instructional portion of the TFL task may have contributed to the pattern of results. For this analysis, we focused on students who scored higher on pretests, but who had different results on the transfer portion of the task. The purpose of this analysis was to generate hypotheses that could be explored in future research.

Findings

Quantitative Findings. For both tasks, we found a statistically significant ($p < .01$) correlation between the Cognition posttest and the overall score on the TFL task. The correlation was higher for the “Fraction Division” task ($r=.68$) than for the “Fraction Generalization” task

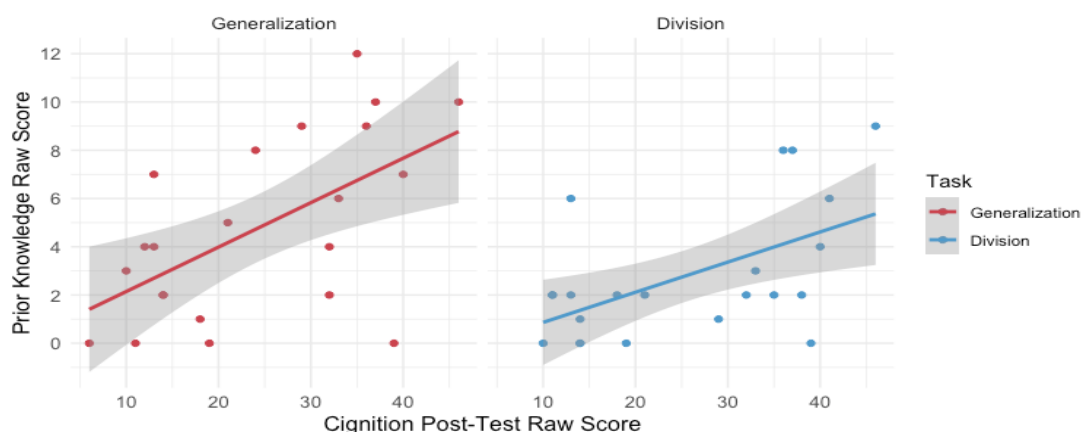
($r=.57$). It is also worth noting that student performance on the Cognition posttest varies widely: students had more or less prior knowledge. It is also worth noting the distribution of transfer scores even for students with higher prior knowledge; transfer was difficult for some students with higher prior knowledge.

Figure 1: Relationship between Cognition Posttest Score and TFL Score
For Generalization: $r = 0.57$, $p < 0.01$ For Division: $r = 0.68$, $p < 0.01$



Additionally, there were statistically significant results between the Cognition posttest and the *prior knowledge* (portion A) part of the task for both “Fraction Division” and “Fraction Generalization” tasks, significance level of $p < 0.01$ for both tasks.

Figure 2: Relationship between Posttest Score and Prior Knowledge (Portion A) of Task
For Generalization, $r = 0.58$, $p < 0.01$ | For Division: $r = 0.54$, $p < 0.01$



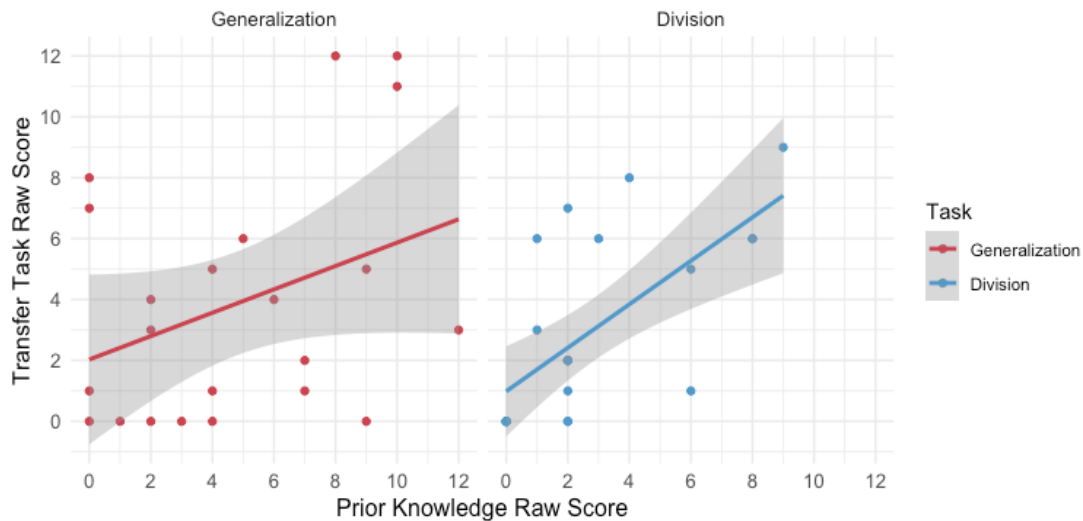
For the relationship between the Cognition posttest and only the *transfer* (portion C) part of the task, the relationship was significant for the “Fraction Division” task, $p < 0.001$ but not significant for the “Fraction Generalization” task.

Figure 3: Relationship between Cognition Posttest Score and Transfer (Portion C) of Task For Generalization: $r = 0.36$, non-significant | For Division: $r = 0.69$, $p < 0.01$



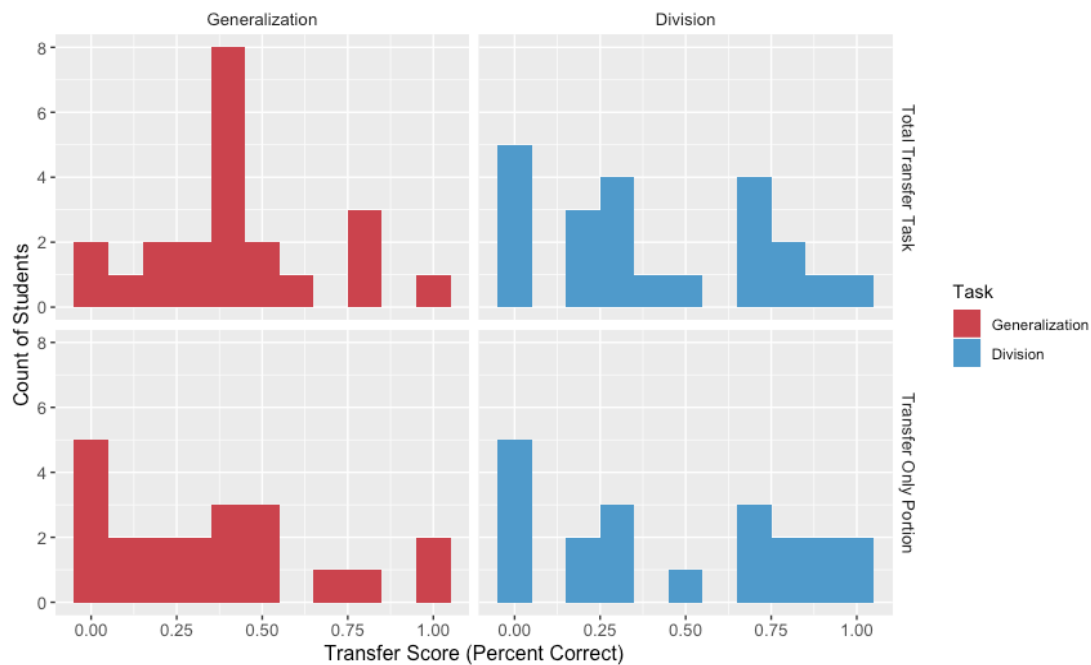
Similarly, the relationship between the *prior knowledge* (portion A) of the task and only the *transfer* (portion C) of the task, the relationship was not significant for the “Fraction Generalization” task but was significant for the “Fraction Division” task, $p < 0.001$.

Figure 4: Relationship between Prior Knowledge and Transfer Parts of Task For Generalization: $r = 0.36$, non-significant | For Division, $r = 0.66$, $p < 0.001$



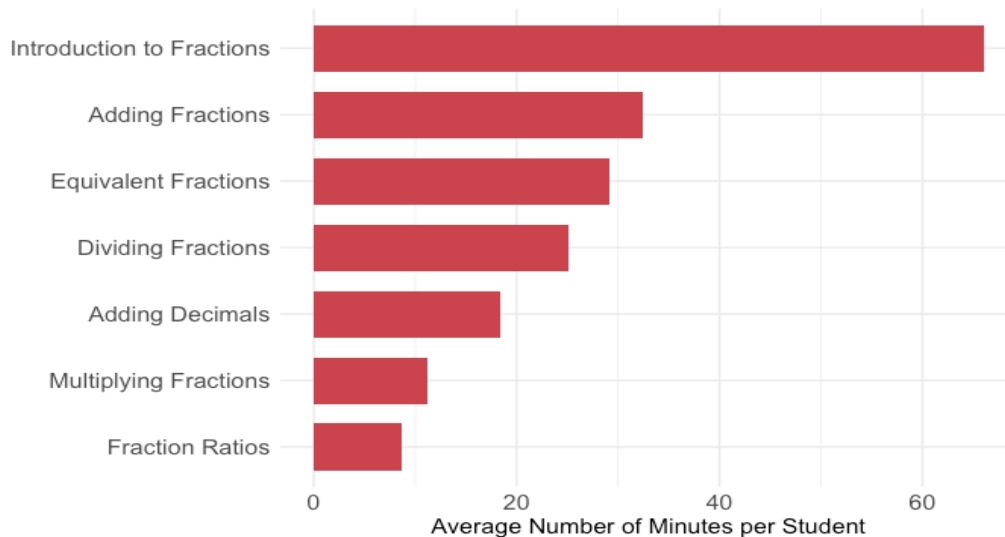
We also looked at the spread of outcomes on the TFL tasks. The histograms below show that most students got 50% or fewer of the available points on the TFL tasks. This is true whether we count both the prior knowledge and transfer portions together or only the transfer portion.

Figure 5: Histograms showing Frequency of Percent Correct
Few students earned > 50%



Regarding the data about tutoring, students completed a total of between 15 and 21 tutoring sessions, with an average of 18.8 sessions per student. With 25-minute long sessions, this equates to a total of nearly 470 tutoring minutes. Regarding the use of the “Fog Stone Isle” game, students played between 0 and 92 sessions (where only one student engaged in 0 game sessions), with an average of about 32 game-sessions per student. These sessions were, on average, about six minutes each, and students attempted about nine problems in each session. Figure 6 shows the average minutes of FSI gameplay in each domain.

Figure 6: Average Minutes of FSI Gameplay in Each Domain



Qualitative Findings. Our team made several observations relevant to why transfer was hard for students. For “Fraction Division,” initial observations revealed transfer was hard for students mostly due to the extent to which students learned to connect multiplicative/additive strategies to that of division. Students often used multiplication or repeated addition to solve problems within the instructional section of the tasks. It was hard for tutors to help students make connections between the operations of multiplication and division so students lacked transfer of one operation to another. This may be due, in part, to each tutors’ own mathematical ability to connect the operations (see Limitations).

For “Fraction Generalization,” transfer was hard given the varied extent to which students learned and used variable notation. As stated above, the instruction section of each TFL task was at the tutor’s discretion, there was no prescribed curricula. As a result, during the instruction portion of the tasks, some tutors focused on numerator/denominator concepts and fraction magnitude, and spent little time on the concept of a variable. In addition, some tutors made the instructional section more tutor-led, so these students engaged with the variable notation less than others. Finally, there were not enough opportunities within the task to assess and/or address the need for understanding that n can represent different values on each side of an expression.

Additionally, observations were made for both tasks regarding the relationships each tutor had with students. Tutors seemed to build both mathematical identity and agency for each student they worked with. Tutors encouraged students to face challenges and continue through difficult problem-solving situations. This seemed to provide students with better capacity to engage with the mathematical problem solving and feel ownership of their accomplishments.

Discussion of Phase 3

Overall, the findings contribute to validating the TFL tasks because the results were consistent with our hypotheses and expectations.

The finding of a statistically significant correlation between the Cognition posttest and our overall TFL task confirms our broad hypothesis: students who have stronger fractions knowledge at the end of the Cognition tutoring ought to be better able to transfer that knowledge to more advanced mathematics. The correlations between the Cognition posttest and the transfer part (portion C) of the Digital Promise tasks offers a more fine-grained view of the same underlying trend; it focuses more tightly on transfer but there were fewer points for students to gain and thus less data underlying the result.

The correlation between the prior knowledge (portion A) and the transfer (portion C) of the TFL tasks is an inner validation of our approach. We expected the prior knowledge section both to elicit prior knowledge and to measure readiness for transfer. If prior knowledge was not displayed here, it would suggest students were not ready for transfer. Consequently, the higher transfer scores for students with higher prior knowledge scores confirms the inner workings of the TFL tasks.

There are many noteworthy limitations to our Phase 3 study. First, due to logistical and recording quality issues beyond our control, we had quite a small data set. This work should be replicated or extended with a larger group of students. Second, we had anticipated testing for transfer with students who were already at or above a threshold for prior knowledge, whereas in the actual data, some students were still showing weaker prior knowledge. The work should be replicated or extended with larger numbers of students who are ready for transfer. Third, we originally wanted for students to have unlimited time for the TFL tasks, but due to logistical issues, had to settle for time-limited tasks. We also observed many factors which could contribute to noisy data. These include challenges with the video conference environment, need for more tutor training, places we could refine the design of the TFL tasks, and like issues.

Conclusion

We conducted three phases of work to design, develop and gather preliminary validation data for a novel approach to assessing Transfer for Future Learning (TFL).

With regard to helping Cognition with product development, the process and findings were helpful in several regards. The early design phase of the work clarified the expected learning progress for Cognition's tutoring and what kinds of tutoring decisions could have later transfer implications. This information was seen as useful for improving the product. In the agile phase (Phase 2), we discovered much about the math content, the tutoring process and the video conferencing software that was informative to Cognition's improvements in these areas. Finally, in as much as Cognition plans to teach division and generalization in more extended tutoring work, the TFL tasks provide useful guidance on how to connect prior knowledge to these advanced and difficult topics.

With regard to informing the assessment field, we took an approach that was grounded in ECD and consistent with a "Preparation for Future Learning" orientation. Specifically, we structured the TFL tasks to determine prior knowledge, to provide instruction to support students' introduction to new content, and then assessed the degree of transfer. Novel features of this work include the forms of domain analysis and domain modeling used to identify concepts that, when well understood by students, can set the stage for transfer of understanding to later topics. Additionally, the design of the instruction portion of the two TFL tasks leverages novel design considerations that can be further explored to inform future TFL task development, including proximity of transfer topics to prior knowledge and task presentation characteristics that cue prior knowledge. While the validation study reported here presents positive findings, more validation is clearly needed as we had access to only a small sample of students. Nonetheless, the field can contemplate the overall approach and the designs tested here as illustrating new directions forward.

Finally, there is more work to do with Cognition's approach to advance our ability to measure transfer. We could better operationalize the degree of transfer in terms of clear criteria for more proximal or more distal tasks. We observed that the tutors were providing social-

emotional support during the TFL tasks, and would like to better understand the role of such support. There is more to do to figure out the best ways to train tutors for mathematical understanding, instruction that supports transfer, and for giving TFL tasks. To understand if the approach is replicable, a next step would be to use the same approach to design more tasks. Overall, there is work to do to conceptualize transfer more fully in this setting, with regard to how a practical measure of transfer can shape formative evaluation and improvement of tutor-based program that is scaling up, and to measure the long-term impacts on student learning.

References

- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological bulletin*, 128(4), 612.
- Booth, J.L., & Newton, K.J. (2012). Fractions: Could they really be the gatekeeper's doorman? *Contemporary Educational Psychology*, 37(4), 247-253.
<https://doi.org/10.1016/j.cedpsych.2012.07.001>
- Bransford, J. D., & Schwartz, D. L. (1999). Rethinking Transfer: A Simple Proposal with Multiple Implications. *Review of Research in Education*, 24, 61–100.
- Chin, D.B., Blair, K.P., Wolf, R.C., Conlin, L.D., Cutumisu, M., Pfaffman, J., & Schwartz, D.L. (2019). Educating and Measuring Choice: A Test of the Transfer of Design Thinking in Problem Solving and Learning, *Journal of the Learning Sciences*, 28(3), 337-380, DOI: 10.1080/10508406.2019.1570933
- DeWolf, M., Bassok, M., & Holyoak, K. J. (2015). From rational numbers to algebra: Separable contributions of decimal magnitude and relational understanding of fractions. *Journal of experimental child psychology*, 133, 72-84.
- Empson S.B., Levi L., & Carpenter T.P. (2011). The algebraic nature of fractions: Developing relational thinking in elementary school. In J. Cai & E. Knuth (Eds.), *Early algebraization. Advances in mathematics education* (pp. 409-428). Springer, Berlin, Heidelberg.
- Evans, J. (1999). Building bridges: Reflections on the problem of transfer of learning in mathematics. *Educational Studies in Mathematics*, 39(1-3), 23-44.
- Hackenberg, A. J. (2013). The fractional knowledge and algebraic reasoning of students with the first multiplicative concept. *Journal of Mathematical Behavior*, 32, 538–563.
- Johnson, H.L., McClintock, E., & Hornbein, P. (2018). Ferris wheels and filling bottles: a case of a student's transfer of covariational reasoning across tasks with different backgrounds and features, *ZDM Mathematics Education*, 49, 851-864.
<https://doi.org/10.1007/s11858-017-0866-4>
- Lobato, J. (2012). The actor-oriented transfer perspective and its contributions to educational research and practice. *Educational Psychologist*, 47(3), 1-16.
- Lobato, J. (2009). Alternative perspectives on the transfer of learning: History, issues, and challenges for future research. *Journal of the Learning Sciences*, 15, 431-449.

- McMullen, J., Laakkonen, E., Hannula-Sormunen, M., & Lehtinen, E. (2015). Modeling the developmental trajectories of rational number concept (s). *Learning and Instruction*, 37, 14-20.
- Mislevy, R., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-Centered Assessment Design*. Educational Testing Service.
- National Research Council. (2000) *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. Washington, DC: The National Academies Press.
<https://doi.org/10.17226/9853>.
- Rayner, V., Bernard, R. M., & Osana, H.P. (2013, April). *A Metaanalysis of transfer of learning in mathematics with a focus on teaching interventions*. Paper presented at the 2013 meeting of the American Educational Research Association, San Francisco, CA.
- Rittle-Johnson, B., Loehr, A. M., & Durkin, K. (2017). Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM: The International Journal on Mathematics Education*, 49(4), 599–611.
- Roschelle, J., Cheng, B., & Cohen, M. (2019). *Report on Expert Panel, Transfer for Future Learning Project*. San Mateo, CA: Digital Promise
- Saxe, G.B., Diakow, R., & Gearhart, M. (2013). Towards curricular coherence in integers and fractions: The efficacy of a lesson sequence that uses the number line as the principal representational context. *ZDM (International Journal on Mathematics Education)*, 45, 343-364.
- Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, 22(2), 129-184.
- Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., ... Chen, M. (2012). Early Predictors of High School Mathematics Achievement. *Psychological Science*, 23(7), 691–697. <https://doi.org/10.1177/0956797612440101>
- Thompson, P. W., & Saldanha, L. (2003). Fractions and multiplicative reasoning. In J. Kilpatrick & G. Martin (Eds.), *Research companion to the NCTM Standards* (pp. 95-113). Washington, DC: National Council of Teachers of Mathematics.
- Tzur, R., Johnson, H. L., McClintock, E., Kenney, R. H., Xin, Y. P., Si, L., . . . Jin, X. (2013). Distinguishing schemes and tasks in children's development of multiplicative reasoning. *PNA*, 7(3), 85-101.

Appendix 1: Mathematica Evaluation Metrics

Mathematica worked with us and with all the other grantees in our cohort. Starting in January 2019 and periodically throughout the project, we met with Mathematica to define a set of metrics to be used in reporting this work to the Bill and Melinda Gates Foundation. The Foundation was one of the three grantors for this work. Over the course of our exploratory research, the nature of the TFL tasks changed considerably. The January 2019 metrics, however, were considered to be “pre-registered” and did not change. Unfortunately, our ability to interpret the pre-registered metrics declined as they became more distant from the evolved TFL tasks. For completeness, we report the pre-registered metrics here. At the end, we reflect on our difficulties in trying to reconcile a pre-registration approach with an iterative validity argument approach.

Expert Panel Review

Three experts, Ann Edwards, Jessica Tsang, and Terry Vendlinks, met with us in Phase 1 of this work to review the quality of our domain analysis and domain modeling for the TFL tasks. In a first metric, we specified these levels of expert rating for our work:

Okay. The expert panel’s consensus is that there is some rationale present in both the domain analysis and domain modeling, but there are substantial and extensive needs for clarification and revision.

Good. The expert panel’s consensus is that there is adequate rationale present in both the domain analysis and domain modeling, but specific clarifications and revisions would yield necessary improvements.

Great. The expert panel’s consensus is that there is strong rationale present in both the domain analysis and domain modeling, with only minor improvements being necessary before proceeding to the next phase.

The expert consensus was that the domain analysis was “good” (sound, but needing clarifications). We decided to make the clarifications and move the project. With regard to domain modeling, we presented three design patterns and accompanying tasks. The experts rated one pattern as “good” and two as “okay.” On this basis, we decided to move forward with only Fraction Division and Fraction Generalization. We incorporated changes to our plans based on expert feedback.

In addition, we noted that the experts were highly enthusiastic about the unique combination of a conceptual game and tutoring that Cognition was developing. They were excited about the transfer work, and the strong potential to develop transfer assessments using a mix of game analytics and human tutors. The main concern of the experts was that the timeline is ambitious (later, we asked for and received a no-cost extension).

Count of Task Revisions

This metric captured how many revisions we made to the TFL tasks after the last round of development (Phase 2) and before using them with a larger number of students (Phase 3). We made one revision to Transfer Division and no revisions to Transfer Generalization. Consequently, one untested revision was present in the Transfer Division task.

We struggle to interpret this metric, as we are aware we could have made more revisions and improvements if we had more time. Consequently, we no longer consider it meaningful to apply the pre-agreed upon judgements of “okay,” “good,” or “great” to the number of revisions. Doing so would obscure the arbitrariness of the metric (given that we could have chosen to do more or fewer revisions).

In the main body of this report, we report that we made a more holistic judgement that the TFL tasks were good enough to move forward with. Likewise, in the main report, we noted sources of possible construct irrelevant variance to be considered as a limitation. We prefer that readers consider these judgements and not the count of revisions.

Task Completion in the Last Round of User Testing

Another metric looked at how many students completed the TFL tasks in the last round of testing. Two of four students (50%) completed all the items in the final user testing on portion A of Fraction Division. One of three students (33%) completed all the items presented in portions A, B, and C of Fraction Generalization. We do not believe these numbers can be meaningfully interpreted.

As the research evolved, the meaning of the “last round of testing” changed considerably, completely changing the interpretation of this metric. Initially, we had assumed that students would complete tasks in a game environment with unlimited time available. In the envisioned game, task completion would indicate that students were able to independently persist to the end of each TFL task. However, as discussed in the main report, we changed the presentation of TFL tasks to occur during tutoring and unlimited time was not available for tutoring. Thus, time available and not student persistence became the limiting factor in task completion. This made the metric much less relevant.

A further problem was that we had initially assumed that in the last round of user testing, we would be testing the entire TFL task from beginning to end. However, our access to tutors and students for testing was limited and consequently in the final round for Fraction Division, we only tested a change to portion A.

Legacy Correctness Metric

In January 2019, we specified a metric based on the idea that students would be playing a game with unlimited time to complete transfer tasks. As with other metrics, this one had specified levels for “okay,” “good,” and “great.”

As mentioned above, by February 2020 we knew that instead students would be doing transfer tasks in the context of an online session with their tutor, with limited time. When time is limited, a metric stated in terms of “all items correct” does not make sense because students may not have time to get to all items.

A further issue is that as the TFL tasks evolved, so did our scoring procedure. “Correctness” was one component of the final scoring procedure, but students also earned points for giving an appropriate, non-trivial explanation (tutors prompted students for explanations). Thus, this January 2019 metric requires us to report and judge a score that is no longer a fit to the design of the assessment tasks. Further, the January metric would register a judgement of below “okay” on the pre-registered ratings, but we argue that this is irrelevant, because the scoring process required by the metric does not fit the assessment task.

	Legacy: Complete whole task correctly	Traditional: Number of items correct
Transfer Generalization	18%	64%
Transfer Division	9%	48%
Overall	14%	56%

Table 2: Metric calculated with legacy scoring procedure yields very low scores, as can be seen by comparison to a more traditional scoring procedure.

We see this data as supporting our concerns about the validity of this legacy metric. For “Complete whole task correctly,” the average student score is 14%, a low score which could indicate the task was much too hard. By way of comparison, a more reasonable scoring procedure which gives credit for correct sub-items shows that students got 56% of the TFL items correct. Neither of these scoring procedures, however, uses the correctness + explanation score, which is discussed in the main report and is the scoring procedure that we determined to be the best fit to the assessment design.

Correlation from Cignition Posttest to TFL Task Scores

Fortunately, the final metric we agreed to with Mathematica did not specify a scoring procedure, so we were able to use our final scoring procedure (which included both correctness and explanation components). Consequently, this metric remains meaningful and interpretable. We present our findings on pages 13 and 14 of the main report. Our finding is that it supports the validity of the TFL tasks. See the discussion in the main report.

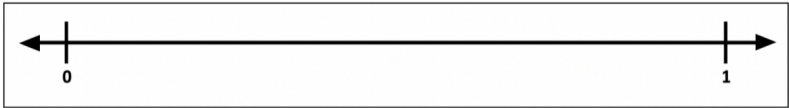
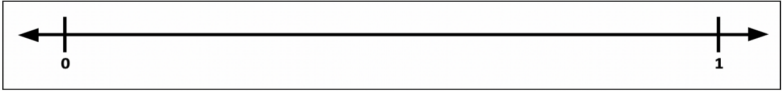
Reflections

Our research and development plan for this transfer assessment project was grounded in Evidence-Centered Design. ECD is grounded in the idea of making an evolving assessment argument as a team proceeds through iterative stages of design and empirical validation of a novel assessment. An assessment argument gathers and presents emerging data on the technical quality of an assessment. The argument documents how issues noted in earlier rounds were resolved in later rounds. The argument is subject to expert critique. Thus, “judgement” or “evaluation” in ECD anticipates a dialectical process of argument and critique among peers. This is the approach we have taken.

The evaluation plan we agreed on with Mathematica for this transfer project focused on pre-registered metrics. In advance, we were asked to create judgements (“okay,” “good,” “great”) for a set of anticipated observables. We were bad at anticipating metrics; we suspect it would be hard for anyone to anticipate metrics for exploratory research. As expected in exploratory research, tasks changed in ways we were not able to anticipate—most notably, in the change from game-based to tutor-based assessment and from untimed to time-limited administration. Consequently, we struggled throughout the project to make sense of most of our pre-registered metrics. One exception was with the last metric (“Correlation from Cognition Posttest to TFL Task Scores”); this is a standard metric in the assessment community, called “concurrent validity” and it remained appropriate.

The ECD and pre-registration-of-metrics approaches were not easy to intermingle. We were frustrated with having to spend considerable time and money on pre-planned data analyses that we knew were going to be hard to interpret. In the future, we would urge care in applying a pre-registered metrics approach to an exploratory assessment development project that takes an iterative validity argument approach.

Appendix 2: Examples from the Fraction Generalization Task

Task Portion	Objective	Example
<p>Portion A: Elicit student's <i>prior knowledge</i></p>	<p>This section is used to determine whether or not students are ready for transfer. It elicits relevant prior knowledge that is necessary for the transfer to take place. For the Fraction Generalization Task, the items are designed to examine students' knowledge of the following:</p> <ul style="list-style-type: none"> • Numerator and denominator concepts • Fraction magnitude • Number line representations <p>The tutor does not provide any instruction or scaffolding in this section, and focuses only on probing for conceptual understanding.</p> <p><i>One item from this section is shown.</i></p>	<p>B) Place $\frac{3}{6}$ and $\frac{4}{6}$ on the number line. Explain which one is greater and why.</p> 
<p>Portion B: <i>Instruction</i> to promote transfer to a new topic</p>	<p>This section is designed to build on students' prior knowledge and extend their understanding to what we call fraction generalizations. Fraction generalizations are generalizations students make and apply to their understanding of fraction concepts. This involves identifying commonalities across fractions/cases and extending reasoning beyond one particular fraction/case. In this section, the tutor begins by introducing and practicing variable notation with the student. The tutor then moves to the number line to compare fractions with like denominators (see example to the right) and like numerators, in an effort to support students as they make fraction generalizations. Some generalizations students make in this section include the following:</p> <ul style="list-style-type: none"> • For any two fractions with the same denominator, the fraction with the smaller numerator will be less than the other fraction. • For any two fractions with the same denominator, the fraction with the larger numerator will be greater than the other fraction. • For any two fractions with the same numerator, the fraction with the smaller denominator will be greater than the other fraction. • For any two fractions with the same numerator, the fraction with the greater denominator will be less than the other fraction. <p>The tutor provides instruction and scaffolding in this section, and tutors can use any approach they see fit to support students.</p>	<p>Pretend we spun the spinner to figure out what number will go in the numerator and it landed on 3.</p> <p>What fraction do we have now? Place it on the number line.</p> <p>[Student places $\frac{3}{6}$ on a blank number line. We assume students can do this correctly.]</p>  <p>Instead of a spinner, let's use n: $\frac{n}{6}$. If you place $\frac{n}{6}$ to the left of $\frac{3}{6}$, is $\frac{n}{6}$ less than or greater than $\frac{3}{6}$? Why?</p> <p>What number can n be if you want to place $\frac{n}{6}$ to the left of $\frac{3}{6}$? Are there any other numbers n can be if $\frac{n}{6}$ is to the left of $\frac{3}{6}$? Explain your thinking.</p> <p>[Tutor can write $\frac{n}{6}$ on whiteboard as talking about it. Student responds. We assume students can do this correctly.]</p> <p>If you place $\frac{n}{6}$ to the right of $\frac{3}{6}$, is $\frac{n}{6}$ less than or greater than $\frac{3}{6}$? Why?</p> <p>What number can n be if you want to place $\frac{n}{6}$ to the right of $\frac{3}{6}$? Are there any other numbers n can be if $\frac{n}{6}$ is to the right of $\frac{3}{6}$? Explain your thinking.</p>

<p>Portion C: Assess performance on the new topic via transfer items</p>	<p>The transfer items reveal to what extent students have built on their prior knowledge and moved toward the target for transfer. For the Fraction Generalization Task, the transfer items are designed to examine students' knowledge of the following:</p> <ul style="list-style-type: none"> • Variable notation • Generalizations related to numerator and denominator concepts • Generalizations related to fraction magnitude <p>The tutor does not provide any instruction or scaffolding in this section, and focuses only on probing for conceptual understanding.</p> <p><i>One item from this section is shown.</i></p>	<p>Explain if each expression below is never, sometimes, or always true for values of n.</p> <p>(A) $n/2 < n/4$</p>
---	--	--